View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by DIAL UCLouvain

Digital access to libraries

"Conventions de transcription régissant les corpus de la banque de données VALIBEL"

Bachy, Sylviane ; Dister, Anne ; Francard, Michel ; Geron, Geneviève ; Giroul, Vincent ; Hambye, Philippe ; Simon, Anne-Catherine ; Wilmet, Régine

Document type : *Document de travail (Working Paper)*

Référence bibliographique

Bachy, Sylviane ; Dister, Anne ; Francard, Michel ; Geron, Geneviève ; Giroul, Vincent ; et. al. *Conventions de transcription régissant les corpus de la banque de données VALIBEL.* (2007) 18 pages

Available at:
http://hdl.handle.net/2078.1/165551

[Downloaded 2019/04/19 at 03:18:25]

CONVENTIONS DE TRANSCRIPTION
RÉGISSANT LES CORPUS DE LA BANQUE DE DONNÉES
VALIBEL

(Version revue en juin 2004 ; mise à jour : 18/04/2007)

Sylviane Bachy, Anne DISTER, Michel FRANCARD, Geneviève GERON, Vincent GIROUL, Philippe HAMBYE, Anne Catherine SIMON, Régine WILMET (ordre alphabétique)
Université catholique de Louvain

1. PRÉLIMINAIRES

Le Centre de recherche VALIBEL, dès la création de sa banque de données textuelles orales, a mis au point une série de conventions de transcription dont les grands principes et les justifications ont été exposés ailleurs (voir Francard & Péronnet 1989 ; Francard 1997). Celles-ci étaient largement convergentes avec celles alors en vigueur dans d'autres centres de recherche, notamment les équipes québécoises animées par S. Poplack, P. Thibault, D. Vincent, ou encore le G.A.R.S.

Ces conventions s'inscrivaient dans la perspective d'une comparaison entre des corpus oraux issus de régions différentes de la francophonie (entre autres, des corpus acadiens et des corpus wallons). En outre, elles prenaient en compte les contraintes techniques imposées alors par l'application des outils informatiques à des corpus textuels[1] (concordanciers, lemmatisateurs, etc.).

En 2004, nous avons jugé utile d'apporter quelques modifications aux conventions initiales et de revoir en conséquence les transcriptions existantes[2]. Ces modifications sont justifiées par les progrès réalisés en matière de traitement informatisé des données textuelles[3], par le souci d'alléger le travail des transcripteurs dans des domaines où la précision de la transcription peut être sujette à caution[4] et par une transcription plus homogène de certains phénomènes, dont la variation empêchait un traitement automatisé[5].

1.1. Principes généraux

Les conventions suivent des principes généraux qui
- respectent l'orthographe conventionnelle ;
- rendent compte des phénomènes liés à l'interaction ;
- sont compatibles avec un traitement informatisé des données ;
- valorisent l'oralité des corpus.

Nous ne transcrivons pas les phénomènes phonétiques (prononciation et intonation), mais nous facilitons l'accès aux données orales. Dans la mesure où le support sonore peut être consulté aisément, à partir des transcriptions ou indépendamment de celles-ci, nous estimons que les transcriptions elles-mêmes peuvent faire l'objet d'une « standardisation » plus importante que dans les cas où le support écrit constitue le seul matériau utilisé pour les analyses de phénomènes lexicaux ou syntaxiques.

La transcription d'un corpus oral requiert beaucoup de temps et d'attention. Elle ne peut se satisfaire d'une seule écoute, même de la part d'un transcripteur expérimenté. Les corpus intégrés dans la banque de données VALIBEL font l'objet d'au moins une vérification systématique par une personne différente du transcripteur.

---

[1] Ainsi, il était nécessaire de lier par un tiret toutes les composantes des lexies complexes que nous souhaitions traiter comme une seule unité ; d'où les graphies *parce-que, tout-à-fait*, etc. À l'inverse, il nous fallait introduire un espace entre deux unités distinctes, d'où les graphies que *dis-◊je, ce jour-◊là, l'◊histoire*, etc.

[2] Voir A. Dister et A.C. Simon. 2007. "La transcription synchronisée des corpus oraux. Un aller-retour entre théorie, méthodologie et traitement informatisé", *Arena Romanistica* 1.

[3] Ce qui entraîne la suppression des tirets ou celle des espaces, voir note précédente.

[4] En particulier dans le domaine des chevauchements, dont les plus complexes se laissent difficilement réduire à une notation linéaire.

[5] Sur ce point, nous n'avons fait que renforcer une option présente dès la création de la banque de données, qui s'appliquait surtout au plan lexical, mais moins à la morphologie.

Ce document présente les conventions de transcription appliquées à l'ensemble de la banque de données VALIBEL. Cela n'exclut pas qu'un chercheur utilise des conventions additionnelles, pour une exploitation particulière, en le spécifiant dans la fiche d'identification de l'enregistrement.
Cependant, nous préconisons l'application des conventions détaillées ici pour constituer une transcription de base utilisable par un ensemble de chercheurs. Tout codage complémentaire se fait de préférence dans une autre version annotée du texte. La transcription de base est stable et archivée dans la banque de données; il est toujours possible de l'aménager dans le cadre d'une publication, etc.

Tout corpus retranscrit est accompagné de fiches d'identification précisant le contexte de l'enregistrement, le profil sociolinguistique des informateurs, de l'enquêteur, etc. (voir § 10).

1.2. Consultation des corpus

Quelques échantillons sont reproduits ici, pour illustrer la mise en œuvre des conventions dans des contextes variés. Nous proposons, via une interface en ligne et à certaines conditions :

- un accès aisé aux données sonores (nos corpus font l'objet d'un alignement transcription/son qui permet d'écouter en temps réel des portions de signal sonore à partir de la transcription) ;
- une transcription prosodique précise pour une partie réduite de nos corpus, qui a fait l'objet d'un alignement phonétique et est accessible sous forme de « prosogrammes[6] » (transcription alignée de la F0 stylisée, de l'intensité et de la durée).

Les chercheurs souhaitant consulter nos corpus peuvent prendre contact avec Anne Dister (anne.dister@uclouvain.be).

1.3. Expertise

La maîtrise des techniques de transcription et d'encodage sur support informatique requiert un apprentissage spécifique. Le Centre VALIBEL met son expertise dans ce domaine au service des chercheurs désireux de travailler sur des corpus oraux francophones, à l'occasion de formations et ou de consultances.

---

[6] Outil de transcription semi-automatisée de l'intonation (F0, durée, intensité), développé par Piet Mertens : http://bach.arts.kuleuven.ac.be/~piet/prosogram/

2. PRÉSENTATION DU TEXTE

2.1. Mise en page

Le corpus se présente sous la forme d'un texte continu (au format .txt), interrompu par les seuls retours à la ligne indiquant qu'un autre locuteur prend la parole[7]. Un tour de parole correspond à l'occupation matérielle du canal de parole par un locuteur ; le tour de parole s'achève lorsqu'un nouveau locuteur prend la parole à son tour.

Le texte ne suit pas les règles de la ponctuation standard de l'écrit (pas de virgule, ni de point), à l'exception du point d'interrogation, utilisé pour indiquer une intonation montante dans une question à forme déclarative (voir § 5.1). Les pauses sont notées au moyen d'un symbole spécifique (voir § 3.1). Les débuts de ligne ne commencent pas par une majuscule.

2.2. Identification des locuteurs

Chaque locuteur est identifié au moyen d'un code individuel (5 lettres suivies d'un chiffre, les 3 premières lettres en minuscules renvoyant au nom du corpus, les 2 lettres suivantes en majuscules renvoyant aux initiales du locuteur ; le chiffre est 1 par défaut, les possibilités de 2 à 9 permettent de distinguer les homonymes). Par convention, on distingue l'enquêteur menant l'entretien avec l'aide d'un questionnaire (entretien semi-directif) par l'utilisation du chiffre 0.
Par contre, si le responsable de l'enregistrement participe à la conversation au même titre que les autres informateurs, il sera désigné par un code normal. On insère une tabulation entre l'identification du locuteur et le début de la transcription de ses paroles.
Exemples:

– accCT0 est l'enquêteur :

accCT0	et / est-ce qu'on peut dire que l'accent est quelque chose qui atteint tout le monde / est-ce que tout le monde a un accent en Belgique
accPJ1	bè d/ d'après ce que d/ j'ai j'ai déjà entendu dire oui

– il n'y a pas d'enquêteur, tous les locuteurs participent au même titre à la conversation :

souVC1	il y a pas beau/ il y a pas b/ beaucoup de magasins
souKC1	ah il y a un terrain de foot pour les enfants à côté / une grande plaine
souGH1	ouais / tu lâches les enfants là et tu
souVC1	ah ouais il y a une grande plaine / c'est vrai
souPH1	tu pars tranquille
souGH1	(rire) (silence)
souKC1	ah mais
souGH1	on doit / on doit être beau il a dit Marc / parce que c'est la photo officielle

Quand le locuteur est non identifié, on note les 3 lettres de code du corpus, suivies de XX9 (accXX9). On précise *souGH1 (?)*, si on croit avoir identifié qu'il s'agit de *souGH1,* sans pour autant en avoir la certitude.

[7] Voir aussi le § 3.2 sur la notation des chevauchements de parole.

2.3. Anonymisation des données

En fonction du consentement de participation obtenu et de l'utilisation qui sera faite des données (accord pour une utilisation en interne ou pour toute publication), il sera nécessaire de procéder à l'anonymisation de l'entrevue.
Bien qu'on parle souvent d'anonymisation, la question légale qui se pose est celle de *l'impossibilité d'identifier des personnes* : l'enjeu est que, sur la base des données recueillies et de leur mode de représentation (transcription par exemple), on ne puisse pas identifier les personnes concernées.
On évitera l'anonymisation sur les données originales (le fichier son original sert de sauvegarde) et on veillera à ce que le son anonymisé ne soit pas plus court que l'original (il s'agit de remplacer les passages à anonymiser par du silence et non de supprimer ceux-ci).
Dans la transcription, on optera pour l'utilisation de pseudonymes à la place des noms réels ou des codes des locuteurs, et une fois qu'un pseudonyme a été attribué à un locuteur, il est archivé de façon à en garder trace et à réutiliser le même pseudonyme pour chaque intervention de ce locuteur.
Le pseudonyme est choisi de manière à conserver les caractéristiques du nom original (nombre de syllabes, origine ethnique, etc.).

3. SYMBOLES TEMPORELS

3.1. Pauses

On distingue trois degrés de pause qui correspondent schématiquement à trois durées relatives :

- la pause brève : une seule barre oblique /
- la pause longue : deux barres obliques //
- le silence : entre parenthèses (silence). Le silence correspond à une pause particulièrement longue

Exemple conHP1 des gens // qui devraient être en contact téléphonique / avec la clientèle c'est une approche / intéressante

La barre oblique est précédée et suivie d'un espace : approche◊/◊intéressante.

À la fin de l'intervention d'un locuteur (lors du passage du tour de parole), nous n'utilisons pas le symbole / ou //. Le retour à la ligne suffit pour indiquer qu'une nouvelle prise de parole a lieu.

Par contre, le silence sera noté à la fin du tour de parole du premier locuteur. Cette indication ne préjuge pas de l'attribution du silence à ce locuteur.

Exemple conHP1 tu viens avec nous? (silence) eh / tu viens avec nous? (silence)
conDA1 oui oui

Les pauses pleines de type [ø] sont notées *euh*.

3.2. Chevauchements

3.2.1. Deux locuteurs parlent en même temps

Deux cas de figures peuvent se présenter :
1. Il y a un chevauchement entre la fin du tour de parole d'un premier locuteur et le début du tour du locuteur suivant. Dans ce cas, le changement de locuteur est indiqué par un retour à la ligne et les symboles |- et -| délimitent respectivement le début et la fin de la portion de parole prononcée simultanément.

Exemple :

conFM1 peut-être que tu |- pourrais venir
conDJ1 il n'en est pas question -| n'y songez pas

2. Le locuteur qui a la parole est chevauché par un deuxième locuteur. Le premier locuteur conserve la parole. Dans ce cas, les deux interventions sont présentées dans le tour de parole du premier locuteur, sans retour à la ligne. La portion de texte en chevauchement se trouve entre les symboles verticaux |- et -|. Le nom du second locuteur est indiqué entre crochets angulaires <> au début de sa prise de parole.

Exemple :

conAB1 il s'en va |- à la fin de chaque année <conDB1> non je crois pas -| pour passer l'hiver en Californie

Dans notre exemple, conAB1 prononce sans s'interrompre l'énoncé "il s'en va à la fin de chaque année pour passer l'hiver en Californie". Sur la partie "à la fin de chaque année", conDB1 dit "non je crois pas". conAB1 garde la parole et termine son tour après le chevauchement.

3. Si, dans un plurilogue, deux locuteurs réagissent exactement en même temps, on utilise la notation suivante qui ne donne la préséance à aucun des deux :

Exemple :

ulaEC1 tu es en quelle année ?
ulaEP1 première licence
ulaEC1 |- première licence
ulaEB1 première licence -|

Les symboles |- et -| sont toujours précédés et suivis d'un espace.

3.2.2. Plus de deux locuteurs parlent en même temps (chevauchements multiples)

Un chevauchement multiple implique plus de deux locuteurs à la fois. On utilise les symboles |-- et --| pour transcrire ces passages.

1. À la frontière de tour. On note les locuteurs qui parlent en chevauchement dans un ordre aléatoire, puisque tous parlent en même temps, en terminant par le locuteur qui conserve la parole.

jeuZM1 elle va |-- raconter
jeuDG1 tu lui dis / et elle doit raconter une histoire
jeuBE1 et moi je dois je dois placer les mots
jeuFJ1 ok // je --| choisis la la |- la sit/
jeuZM1 oui -| |- tu <jeuBE1> (xx) -| choisis

irtAA1 Dominique nous appelle de Corbeil
irtXD1 oui bonjour
irtKA1 |-- bonjour
irtAA1 bonjour
irtXD1 bonjour madame bonjour --| messieurs hum

2. Dans un chevauchement interne, lorsque le locuteur qui a la parole au début du chevauchement conserve la parole après que le chevauchement a eu lieu:

Exemple :

conLK1 et c'est comment |-- visu/ visuel hein visuel <conGG0> hein ouais ouais ouais pour retrouver <conWR1> un mémo là / pour retrouver euh / oui --| c'est incroyable

3.2.3. Les pseudo-tours de parole

Notre définition du tour de parole étant liée à l'occupation matérielle du canal par un locuteur, les productions de type régulateur comme *m* qui interviennent durant une pause du locuteur principal ne constituent donc pas un chevauchement et doivent être transcrites dans un tour de parole à part entière.

exeDF1 je me levais à cinq heures / et
exeDA1 mm
exeDF1 je prenais le premier tram pour aller à la mine

Si le régulateur est prononcé en chevauchement, il est transcrit à l'intérieur du tour de parole en cours.

exeDF1 je me levais à cinq heures / |- et <exeDA1> mm -| je prenais le premier tram pour aller à la mine

3.2.4. Notation des chevauchements et coupure de mots composés

Lorsque le chevauchement ne concerne qu'une partie de mot, on considérera que le mot partiellement chevauché appartient totalement à la séquence chevauchée. On ne coupera donc pas celui-ci dans la graphie. Dans l'exemple suivant, seul le début du mot *syllabes* est prononcé en même temps que la fin de l'énoncé de accCT0. On transcrit néanmoins *syllabes* totalement dans le chevauchement de parole.

Exemple :

accBF1 on me l'a dit et parfois je m'en rends compte aussi / |- je traîne sur des syllabes <accCT0> tu t'en rends compte -| qu'on m'a déjà dit

Afin d'éviter des divergences dans les transcriptions dues à des choix théoriques différents, nous adoptons ici une définition typographique de la notion de mot. Un mot est une suite de lettres entre deux séparateurs (espace, trait d'union, ponctuation). On considérera donc qu'il y a 3 mots (graphiques) dans *pomme de terre* et 2 mots (graphiques) dans *grand-mère*.

Si seulement l'un de ces mots est chevauché dans l'énoncé oral, lui seul sera inclus dans la portion de la transcription notant le chevauchement.

Exemple :

exeAD0 c'est fou ce que le prix de la pomme |- de terre a augmenté <exeBD1> je sais -| cette année

Si le séparateur est un trait d'union, celui-ci se rattachera directement au mot graphique qui le précède.

Exemple :

exeED1 oui c'est comme |- ma grand- <exeBD1> je sais -| mère qui disait ça
exeFA0 mais donne- |- moi le pain
exeJH2 demande à -| Pierre

3.3. Séquences simultanées (conversations en parallèle)

Dans le cas où au moins quatre locuteurs participent à une conversation, il peut arriver que celle-ci se scinde en plusieurs échanges distincts (cf. Sacks *et al.* 1974). Les paroles d'un locuteur donné peuvent alors être produites en même temps que celles d'un autre locuteur, sans pour autant que les deux conversations soient liées. Ces conversations différentes coïncident donc temporellement, mais sans se situer dans le même cadre interactionnel.

Lorsque le transcripteur a la certitude que deux séquences de parole concomitantes ne prennent pas place dans le même cadre interactionnel, il considérera qu'il ne s'agit pas d'un fait de chevauchent (cf. supra) mais bien de *séquences simultanées*[8].

Pour noter ces séquences simultanées, on procédera comme suit. Au début de l'intervention pendant laquelle une séquence simultanée apparaît, on notera le signe |§ suivi d'un retour à la ligne. On transcrira ensuite toutes les interventions d'autres locuteurs (avec les chevauchements éventuels) qui prennent place dans le même cadre interactionnel et qui ont lieu (ne fût-ce que partiellement) en même temps que l'autre séquence. On indiquera ensuite le signe § suivi et précédé d'un retour à la ligne et l'on notera à la suite les interventions successives qui ont lieu simultanément dans un second cadre interactionnel. Il en va de même si un troisième cadre coïncide. À la fin de la dernière intervention de la séquence qui a lieu dans le même intervalle temporel, on notera le signe §| (précédé et suivi d'un retour à la ligne) afin de marquer la fin de l'intervalle durant lequel on observe des séquences simultanées.

Exemple (les lignes sont numérotées pour la clarté du propos)

1 exeJA1 je vais vous demander de travailler par groupes de deux et de le faire
2 |§
3 dans le silence en essayant de repérer les passages où l'auteur insiste sur les signes de
4 l'agonie du héros
5 §
6 exeLA1 j'en ai marre de toujours faire ça
7 exeMO1 c'est exactement la même chose que la semaine |- dernière
8 exePR1 pas -| tout à fait
9 §
10 exeME1 (xxx) pas recommencer avec ça
11 exeZP1 tu l'as dit
12 §|
13 exeJA1 vous voulez bien écouter les consignes ?

Dans le cas de ces séquences parallèles, on sera attentif au fait que cette *convention* de transcription rompt la succession temporelle du discours en regroupant en blocs distincts des séquences considérées comme homogènes du point de vue interactionnel. On privilégie donc ainsi la logique interactionnelle par rapport à la logique temporelle, qui voudrait que les paroles prononcées soient transcrites dans leur

[8] Le chevauchement implique donc une interaction entre les séquences qui va plus loin que leur simple coïncidence. Notons néanmoins qu'une séquence de conversation simultanée peut finir par rejoindre une conversation qui lui était simplement parallèle et constituer dès cet instant un chevauchement.

stricte succession chronologique. Ce n'est pas le cas ici, où la transcription réorganise en quelque sorte le discours : « tu l'as dit » (ligne 11) est en fait prononcé avant « c'est exactement la même chose » (ligne 6), pourtant antérieur dans la transcription. Notons que cette transcription s'avère très lisible ; elle a également l'avantage d'éviter la surcharge de didascalies qui, dans une transcription qui respecterait strictement la séquentialité du discours, seraient nécessaires à la compréhension (par ex. « merLA1 s'adressant à merMO1 »).

Il faut encore noter

- qu'il est possible qu'un seul locuteur intervienne dans une séquence parallèle, le locuteur auquel il s'adresse pouvant, par exemple, ne pas intervenir ;
- que si des séquences parallèles débutent pendant le tour de parole d'un locuteur, la fin des paroles du tour de ce locuteur sont transcrites en début de ligne (après l'insertion du signe de début de séquences simultanées), sans que soit repris son code locuteur. Nous avons fait ce choix de transcription afin de ne pas rompre arbitrairement son tour de parole (cf. notre exemple, lignes 1, 2 et 3).

4. AMORCES DE MORPHÈMES

Dans le cas d'une interruption (avec ou sans reprise ultérieure), le mot amorcé est immédiatement suivi de la barre oblique (sans espace).

Exemple: main/ maintenant

Lorsque la reprise constitue la fin du mot amorcé, nous notons : *main/ -tenant* (avec un espace entre la barre oblique et le tiret).

Une amorce de mot en [k] est notée *c/* ou *qu/* ou *k/* en fonction de l'orthographe du mot effectivement réalisé ; de même, une amorce en [s] sera notée *s/*, *c/* ou *ç/*.

Exemples: je l'ai s/ su seulement par après
c/ c'était le problème
ç/ ça aussi il a oublié de nous le dire
qu/ quelle heure est-il
je voulais c/ comprendre

De même pour un cas comme *ils sont venus halbé/ habiter là*, nous proposons une notation analysant toujours la séquence tronquée comme une amorce du lexème qui suit. Quand cette analyse est impossible, on note q/ pour une amorce en [k] et s/ pour une amorce en [s].

Le pronom personnel *il* peut être réalisé variablement comme [il] ou [i]. Dans ce dernier cas, et même lorsqu'il est répété, on transcrit *il* et on n'interprète pas la répétition du morphème comme impliquant la réalisation d'une amorce, ce qui se transcrirait *i/*.

Exemple: il il il pense (prononcé [i i i pa~s])

5. ASPECTS PARAVERBAUX

5.1. Question à forme déclarative avec intonation montante

Le seul emprunt à la ponctuation traditionnelle est celui du point d'interrogation. Mais sa valeur, dans nos corpus oraux, est d'indiquer, dans une construction déclarative, l'intonation montante caractéristique d'une question.
Exemples:

conDA1 tu vas venir ? // qu'il dit
conSA1 vraiment ?

Nous ne transcrivons pas l'intonation dans la transcription orthographique.

5.2. Éléments para-verbaux

Des précisions sur le contexte situationnel (explication d'un bruit; note sur la gestuelle du locuteur) et, plus généralement, tout renseignement nécessaire à la compréhension de la séquence enregistrée, seront notés entre parenthèses. L'apparition de ce type de parenthèse n'implique pas un retour à la ligne.

5.3. Exemples de commentaires (didascalies)

(rire)	(bâillement)	(chuchoté)
(silence)	(inspiration)	(bruit)
(toux)	(chantonné)	(imitation)
(soupir)		

6. UTILISATION DE L'ORTHOGRAPHE STANDARD ET VARIANTES NON STANDARD

Toute forme lexicale non standard qui n'est répertoriée dans aucun dictionnaire du français est orthographiée de manière à refléter la prononciation (voir les exemples sous « lexèmes »). Toute forme qui appartient au français standard mais fait l'objet d'une prononciation particulière (ressentie comme marquée par le transcripteur) est orthographiée de manière standard, et la prononciation est mentionnée en alphabet phonétique SAMPA[9] (entre crochets droits après le mot graphique).

6.1. Prononciations « non standard »

De manière générale, on utilisera le SAMPA pour désambiguïser les prononciations qui divergent considérablement de la forme graphique du mot ou pour souligner une prononciation jugée remarquable ou intéressante. Cela peut concerner une liaison erratique, la prononciation régionale, stylistiquement marquée ou idiosyncrasique d'un mot, etc.

---

[9] SAMPA pour Speech Assessment Methods Phonetic Alphabet*:* http://www.phon.ucl.ac.uk/home/sampa/french.htm. Nous avons abandonné l'API parce que cette police n'est pas compatible avec le format .txt des documents.

Exemples:

je vais toujours voir le football [fOtbal] avec le voisin
il m'a fait une scène [se~n]
il faut qu'il soit [swaj] là
ça m'a coûté cent euros [sa~z2RO]
il a eu un infarctus [e~fRaktys]
on va manger des courgettes [guRZEt]
il a dégringolé [dedRe~gOle] dans les escaliers
c'est une faute de construction [kRo~stRyksjo~]

Nous ne distinguons pas, dans la transcription, les différentes prononciations de *il y a*, en trois, deux ou une seule syllabe(s) : [ilia] [ilja] [ija] [ja]. Quelle que soit la prononciation, nous notons toujours la forme standard *il y a*. Une recherche approfondie sur la prononciation effective de ces formes pourra se faire à partir des corpus alignés (texte/son). Nous ne distinguons pas non plus les différentes prononciations de *il* : [i] ou [il].

Lorsque la chute de certains segments est **prédictible** (p.e. lorsqu'elle suit des règles sandhi dans une variété linguistique donnée), on a choisi de ne pas l'indiquer par un signe graphique (comme l'apostrophe).

Exemples (en français parlé en Wallonie, ci-après : FW) :

prendre pour soi est à « lire » [pRa~t] ou [pRa~dR];
la *fenêtre* du salon est, phonétiquement, [fnEt] ou [f@nEtR]

À la différence d'autres corpus oraux existants, on n'utilisera pas des graphies comme *fenêt'* ou *prent'* (ou *fenête, prende*). L'apostrophe servira par contre à noter les cas d'élision présents à l'écrit :

Exemple :

conWR1	j'aurais demandé s'il pouvait garder Olivier
conFM1	je t'apprends cette nouvelle

On n'élide donc pas *t'aurais* ou *celui qu'est arrivé*.

6.2. Lexèmes non répertoriés

Les lexèmes non répertoriés dans les dictionnaires apparaîtront donc dans une graphie qui respecte l'orthographe française standard et est proche de leur prononciation, comme "appelable", "afonner" et "suggération" dans les exemples suivants

Exemple

conDA1	(parce que nous) il faut bah avec maman qui n'est pas très bien il faut absolument qu'on soit euh / qu'on soit appelable donc il / il faut absolument que euh //
conGV1	il a dû afonner deux pintes
conGG1	il a fait une suggération intéressante à la réunion d'hier

Les particules de l'oral (onomatopées, ponctuants, particules discursives, etc.) sont transcrites de manière uniforme.

6.2.1. Interjections et onomatopées

ah		**m'enfin**	
aïe		oh	
bah		**oh la la, oh là là**	
bè	[bE]	ok	
ben	[be~]	ouais	
eh	[E]	ouille	
enfin	[a~fe~, fe~]	**oula, ouh là**	
euh		p	[p]
gnagnagna		pf	[pf]
hein		pff	[pf:]
hum		**pt**	
m	[m]	**t**	
mm	[mm]	wouf, waf	
moui			
mouais			

6.2.2. Abréviations usuelles

On n'utilise aucune abréviation graphique usuelle finissant par un point. On ne note par *etc*. mais *etcetera*.

6.3. Emprunts

Les mots identifiés comme <u>emprunts</u> seront transcrits suivant les standards orthographiques de la langue source. Pour les langues régionales de Wallonie, on utilise l'orthographe Feller.

Exemples :

> digLG1 et il serait parlé par qui alors / qui est-ce qui continuerait à parler wallon comme ça // dins dins les pus djône qwè hein
>
> digFM0 vive nos' cok ah oui (rire) ben juste c' est la c'est ma djèrinne ma djèrinne question (rires) di où d'où ç'ki vos vos sintoz par exemple vos vos sintoz d'Lutrebwâs bin sûr ça c'est sûr / mès pus grand k'ça c'èst c'èst qwè / vos vos sintoz d' l'Ardène vos vos sintoz do Lucsamboûr vos vos sintoz do' l'Walonie vos vos sintoz do l'Belgique di l'Europe enfin ça
>
> conFM1 <u>il</u> ont
> [emprunt du pronom *il* « ils » au wallon]

Lorsqu'un mot (ou une expression) peut appartenir, sur la base de sa seule forme graphique, à l'une ou l'autre langue, sa <u>prononciation effective</u> par le locuteur est explicitée grâce à la transcription phonétique).

Exemple: norBB1 tout ça c'était top secret [tOpsikRit]

On n'utilise pas d'italiques pour transcrire les emprunts ou les mots non attestés dans les dictionnaires de référence, afin de limiter la part d'analyse au niveau de la transcription des données.

Voici quelques formes régionales et leur transcription:
asteure (régional)
aussi non
nom-di-Djo! (et variantes)

6.4. Morphologie

Accord du participe passé. Lorsque, dans l'accord du participe passé, la marque du féminin n'a pas été entendue (alors qu'elle était attendue), on note la forme non accordée au féminin, mais on note cependant l'accord du pluriel s'il y a lieu, comme nous l'avons fait pour "compris" et "écrits" dans les exemples suivants

Exemples :
conSA1 elle n'a pas été bien compris
conHP1 les lettres que j'ai écrits

Le principe est le suivant : sauf preuve audible d'une absence d'accord (au féminin), on postule que l'accord a été fait correctement par le locuteur.

Particule de négation *n(e)*. Lorsque la présence ou l'absence de la particule de négation *n'* n'est pas audible (entre le pronom *on* et un verbe à initiale vocalique), on note le *n'* d'office.

Exemple : conDA1 on n'est pas certains d'arriver à temps

Variantes verbales non standard. Lorsqu'une variante non standard est créée par un locuteur, le plus souvent sur le modèle d'une forme existant dans la conjugaison régulière, on note la forme telle qu'elle a été réalisée, comme pour *prendu* et *disez* dans l'exemple suivant.

Exemple : j'ai prendu un livre ; vous disez.

7. TYPOGRAPHIE

7.1. Noms propres

La majuscule est utilisée pour marquer les noms propres ou les noms à référent unique : *l'État; la Wallonie*, *Pierre Denis*.

On utilise également la majuscule et le trait d'union dans les titres (de journal, de roman, etc.) :
Exemple : *Le-Soir* (journal), *La-Chartreuse-de-Parme* (roman), *Au-bon-Coin* (café), *À-l'-Ombre-des-jeunes-filles-en-fleurs*

7.2. Chiffres

Les numéraux cardinaux sont écrits en toutes lettres (on distinguera ainsi septante de soixante-dix) et sont unis par des traits d'union :

Exemples : trois-cent-quarante-six-mille-deux-cent-dix-sept

Les fractions sont également écrites en toutes lettres, avec trait d'union :
Exemples : trois-quarts; deux-tiers

7.3. Sigles et acronymes

Les sigles sont transcrits en capitales sans points : FNRS, SNCB. De même lorsqu'un mot est épelé, on note les lettres capitales :

conAB1 oui / et puis |- et puis j'ai arrêté / <conDB1> (mais comment ça s'écrit) cette chose -| T A E |- / K
conDB1 avec deux E -| mm / |- oui
conAB1 W -| O N D O / oui c'est ça / euh

Lorsqu'il s'agit d'un acronyme, on note une majuscule au début et le reste en bas de casse : *Valibel, Setca*, etc.
Ceci permet de distinguer deux prononciations de l'abréviation PUF : PUF [pe.y.Ef] et Puf [pyf].

8. PASSAGES DIFFICILES À TRANSCRIRE

8.1. Multi-transcriptions

On propose une multi-transcription lorsqu'on ne peut pas décider, à l'écoute, de la forme qui a été prononcée. Dans ce cas, on note entre accolades les deux formes possibles, séparées par une virgule (on note un espace *avant* et *après* la virgule).

Exemple : conDJ1 j'ai fait commander {des , deux} cartouches d'encre et des timbres

conCM1 j'essaie de grouper / au maximum mes {conditions , commissions} maman va me regarder de travers parce que / elle dira tu oublies toujours quelque chose

conDM1 {il travaille , ils travaillent}

Enfin, quand le transcripteur doute de sa transcription sans avoir d'autre solution à proposer, il note le segment sur lequel porte l'incertitude entre accolades, sans proposer d'alternative :

Exemple : conCB1 c'était vraiment {trop} // enfin je trouve

Si l'incertitude du transcripteur provient d'un problème de référent, et non d'écoute, il peut choisir la solution qui lui paraît la plus plausible dans le contexte et ne note pas de multi-transcriptions (par ex. il note *j'ai mal aux pieds* ou *j'ai mal au pied*).

8.2. Passages inaudibles

Les parenthèses sont utilisées pour l'indication d'un passage incompréhensible, transcrit (x) (= une syllabe) ou (xx) (= un groupe de syllabes) ou (xxx) (= un passage plus long).
Pour les cas où le transcripteur peut proposer une solution phonétique (mais pas de solution orthographique), cette notation sera suivie de la transcription de la séquence en SAMPA (entre crochets). En aucun cas la notation phonétique ne remplace la transcription graphique.

9. ALTERNANCE DE CODES

Dans certaines transcription plus anciennes, l'alternance entre deux codes (langues, variétés de langues) est notée au moyen de balises ouvrantes (A+) et fermantes (A-). On détermine, pour l'interaction transcrite, une langue de base (en général, le français) et on signale les séquences dans l'autre langue (dans l'exemple ci-dessous, il s'agit de l'espagnol).

Exemple :

digVM0 mais / (A+) el norte de España es // muy diferente a la (A-) Méditerranée

Étant donné que la transcription de l'alternance codique demande une analyse de la part du transcripteur, cette convention a été appliquée à certains corpus qui ont fait l'objet d'une exploitation dans ce domaine. D'autres corpus présentent donc de l'alternance de code (par ex. wallon/français) sans que celle-ci soit signalée explicitement, et on recommande d'ailleurs, dans un premier temps, de ne pas transcrire l'alternance en utilisant ces conventions, mais de laisser le traitement de ce phénomène à une phase ultérieure d'analyse des données.

10. MÉTADONNÉES: IDENTIFICATION ET RÉFÉRENCIATION DES CORPUS

Le logiciel [moca] est utilisé pour l'archivage des métadonnées.

10.1. Fiches d'identification

Chaque entrevue est accompagnée de trois fiches d'identifications concernant :

- le(s) locuteur(s) ;
- l'enquêteur (s'il s'agit d'une entrevue guidée) ;
- l'entrevue (description du contexte, des liens interactionnels entre les interlocuteurs et précisions sur la transcription).

Chaque corpus (regroupant plusieurs entrevues) est également décrit dans une fiche.

10.2. Citation d'un extrait de corpus

Il est important d'accompagner les citations de corpus oraux de références précises, à l'instar de ce qui se pratique pour l'écrit. Les conventions suivantes sont adoptées.

Exemple :

[VALIBEL, 1989, F, employée, 37 ans, Bastogne (Luxembourg), conMM1r]

VALIBEL
[appellation générique pour l'ensemble des corpus oraux disponibles]

1989
[date de constitution (enregistrement) du corpus]

H / F
[sexe de l'informateur]

étudiant, ***agriculteur***, etc.
[catégorie socioprofessionnelle de l'informateur au moment de l'enregistrement du corpus]

37 ans
[âge de l'informateur au moment de l'enregistrement du corpus]

Charleroi (Hainaut)
[origine géographique de l'informateur - première mention : localité; deuxième mention : province. En cas de divergence entre l'origine géographique et la résidence de l'informateur, on choisira la donnée la plus significative (notamment en terme de durée)]

11. ALPHABET PHONÉTIQUE SAMPA POUR LE FRANÇAIS

Présentation tirée de : http://www.phon.ucl.ac.uk/home/sampa/french.htm (au 2.7.2004)

Les Consonnes
Le système consonantique français est composé 20 consonnes réparties en 3 catégories : les occlusives, les constrictives et les semi-consonnes.

Les occlusives ou plosives orales (p b t d k g) et les occlusives nasales (m n J) sont produites par une fermeture total du canal (catastase) vocal suivi d'un relâchement brusque (métastase).
Les constrictives ou fricatives (*f v s z S Z*) et les liquides (*l R*) sont produites par un rétrécissement du conduit vocal en un endroit précis.
Les semi-consonnes ou semi-voyelles (*w H j*) sont produites comme les constrictives mais le lieu de constriction est moins marqué ce qui fait penser au canal libre des voyelles.

Les occlusives peuvent être sourdes (non-voisées) *p t k* ou sonores (voisées) *b d g*.

Symbole	**mot**	**Transcription**
p	pont	po~
b	bon	bo~
t	temps	ta~
d	dans	da~
k	quand	ka~
g	gant	ga~

Il y a 3 **nasales sonores** *m n J* produites par l'abaissement du voile du palais qui permet un passage de l'air par le nez lors de l'occlusion. Une quatrième nasale *N* est seulement présente dans les mots anglais utilisés en français.

m	mont	mo~
n	nom	no~
J	oignon	oJo~
N	camping	ka~piN

Les fricatives peuvent être sourdes *f, s, S* ou sonores *v z Z* .

f	femme	fam
v	vent	va~
s	sans	sa~
z	zone	zon
S	champ	Sa~
Z	gens	Za~

Il y a **deux liquides** *l R* ainsi que 3 **semi-voyelles**. *w H j*.

l	long	lo~
R	rond	Ro~
w	coin	kwe~
H	huit	Hit
j	ion	jo~

Les voyelles
Le système vocalique comprend 16 voyelles: 12 voyelles orales. *i e E a A O o u y 2 9 @,* et 4 voyelles nasales *e~ a~ o~ 9~*.

i	si	si
e	ses	se
E	seize	sEz
a	patte	pat
A	pâte	pAt
O	comme	kOm
o	gros	gRo
u	doux	du
y	du	dy
2	deux	d2
9	neuf	n9f
@	justement	Zyst@ma~
e~	vin	ve~
a~	vent	va~
o~	bon	bo~
9~	brun	bR9~

SAMPA UCL Phonetics and Linguistics University College London. J.C. Wells.

Traduction Sylviane Bachy 14-06-2006